Protein databases

Henrik Nielsen

Background- Nucleotide databases

GenBank, http://www.ncbi.nlm.nih.gov/Genbank/

National Center for Biotechnology Information (NCBI), National Library of Medicine (NLM), National Institutes of Health (NIH), USA.

EMBL, http://www.ebi.ac.uk/embl/

European Bioinformatics Institute (EBI), England

(Established in 1980 by the European Molecular Biology Laboratory, Heidelberg, Tyskland)

DDBJ, http://www.ddbj.nig.ac.jp/

National Institue of Genetics, Japan

Together they form

International Nucleotide Sequence Database Collaboration, http://www.insdc.org/

Protein databases

Swiss-Prot, http://www.expasy.org/sprot/

Established in 1986 in Swizerland

ExPASy (Expert Protein Analysis System)

Swiss Institute of Bioinformatics (SIB) and European Bioinformatics Institute (EBI)

PIR, http://pir.georgetown.edu/

Established in 1984

National Biomedical Research Foundation, Georgetown University, USA

In 2002 merged into:

UniProt, http://www.uniprot.org/

A collaboration between SIB, EBI and Georgetown University.

UniProt

UniProt Knowledgebase (UniProtKB)

UniProt Reference Clusters (UniRef)

UniProt Archive (UniParc)

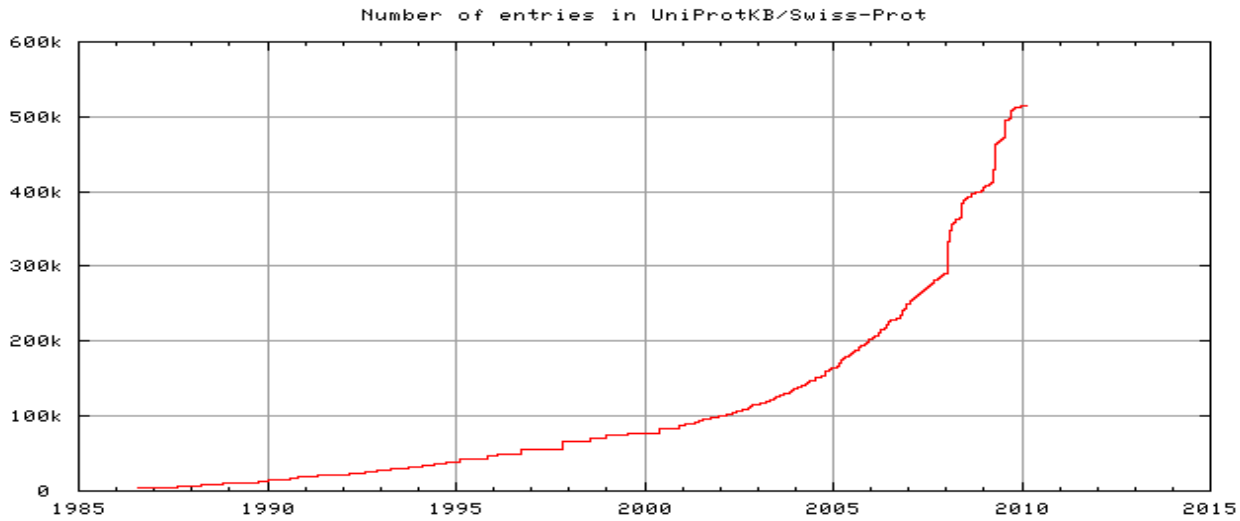
UniProt Knowledgebase Release 2011_02 (08-Feb-11)
consists of:

UniProtKB/Swiss-Prot: Annotated manually (*curated*)
525,207 entries

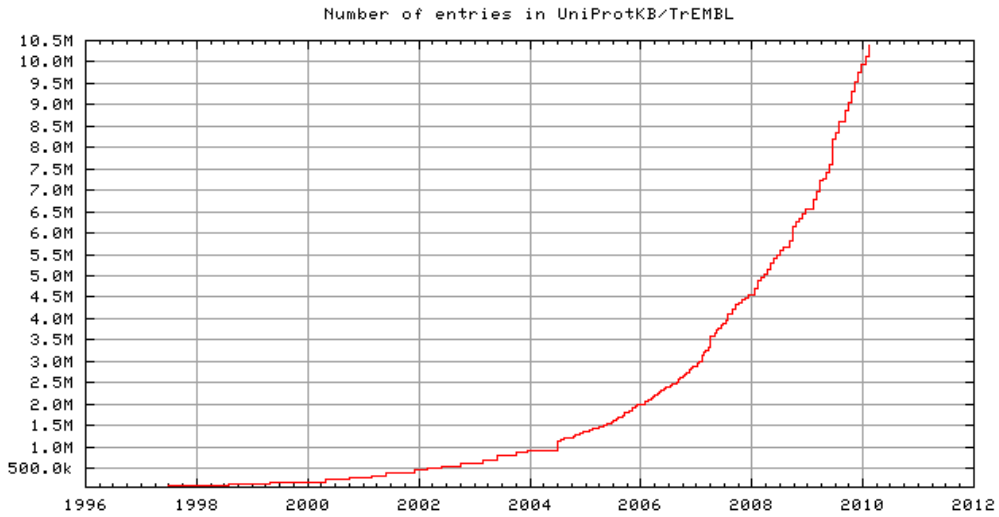
UniProtKB/TrEMBL: Computer annotated
13,499,622 entries

Growth of UniProt

Swiss-Prot



TrEMBL



Content of UniProt Knowledgebase

- Amino acid sequences
- Functional and structural annotations
  - Function / activity
  - Secondary structure
  - Subcellular location
  - Mutations, phenotypes
  - Post-translational modifications
- Origin
  - organism: Species, subspecies; classification
  - tissue
- References
- Cross references

Amino acid sequences

From where do you get amino acid sequences?

- Translation of nucleotide sequences (GenBank/EMBL/DDBJ)
- Amino acids sequencing: *Edman degradation*
- Mass spectrometry
- 3D-structures

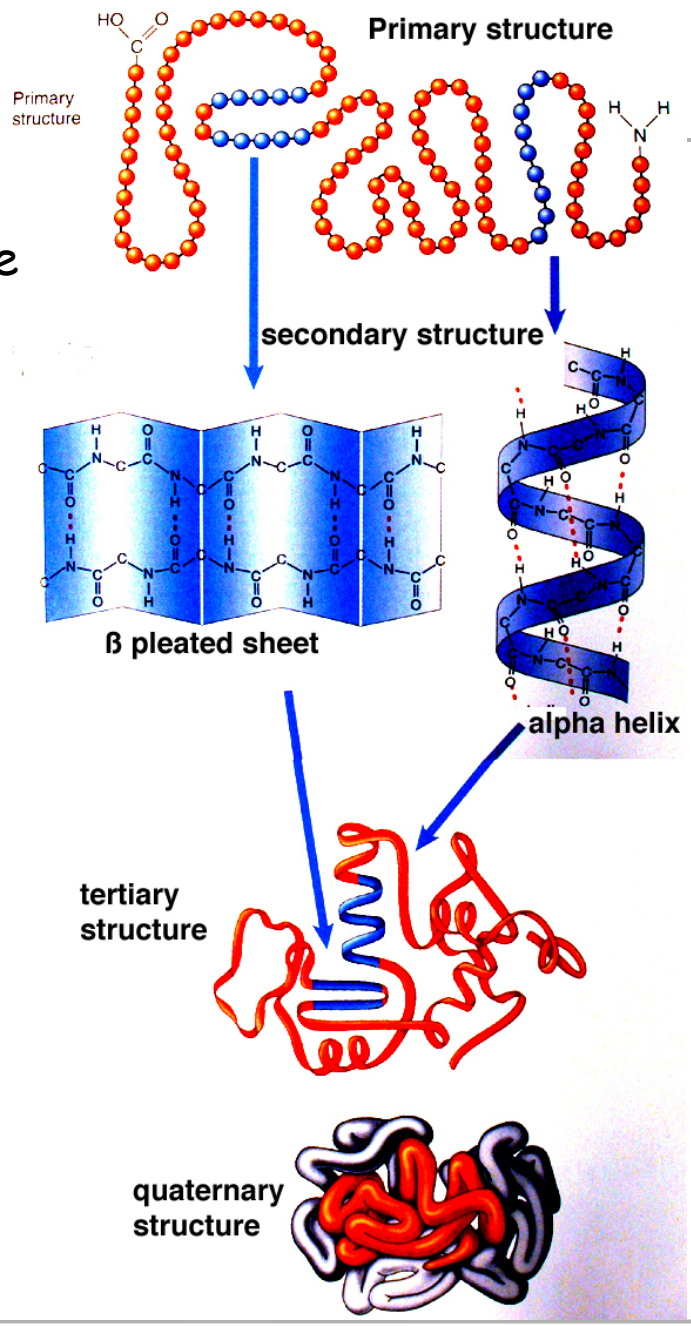
Protein structure

Primary structure: Amino acid sequence

Secondary structure:
"Backbone" hydrogen bonding
Alpha helix / Beta sheet / Turn

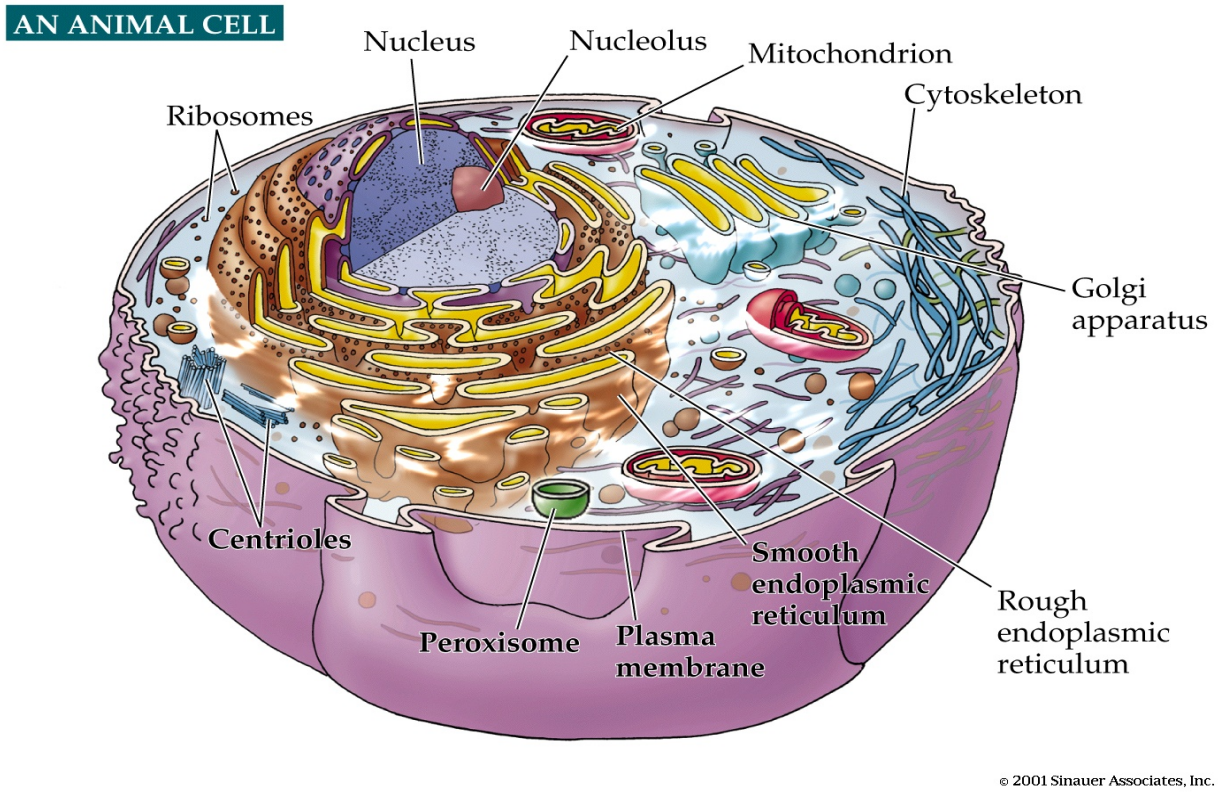
Tertiary structure: Fold, 3D coordinates

Quaternary structure: subunits



Subcellular location

An animal cell:



Post-translational modifications

- Cleavage of signal peptide, transit peptide or pro-peptide
- Phosphorylation
- Glycosylation
- Lipid anchors
- Disulfide bond
- Prosthetic groups (*e.g.* metal ions)

Content of UniProt Knowledgebase

http://www.uniprot.org/uniprot/Q9ULV8

- Name, entry data etc
- Organism
- Functional annotations (comments)
- Sequence
- References
- Cross references
  - 3D structure - PDB
  - EMBL
  - ..

Evidence

3 types of *non-experimental qualifiers* in Sequence annotation and General comment:

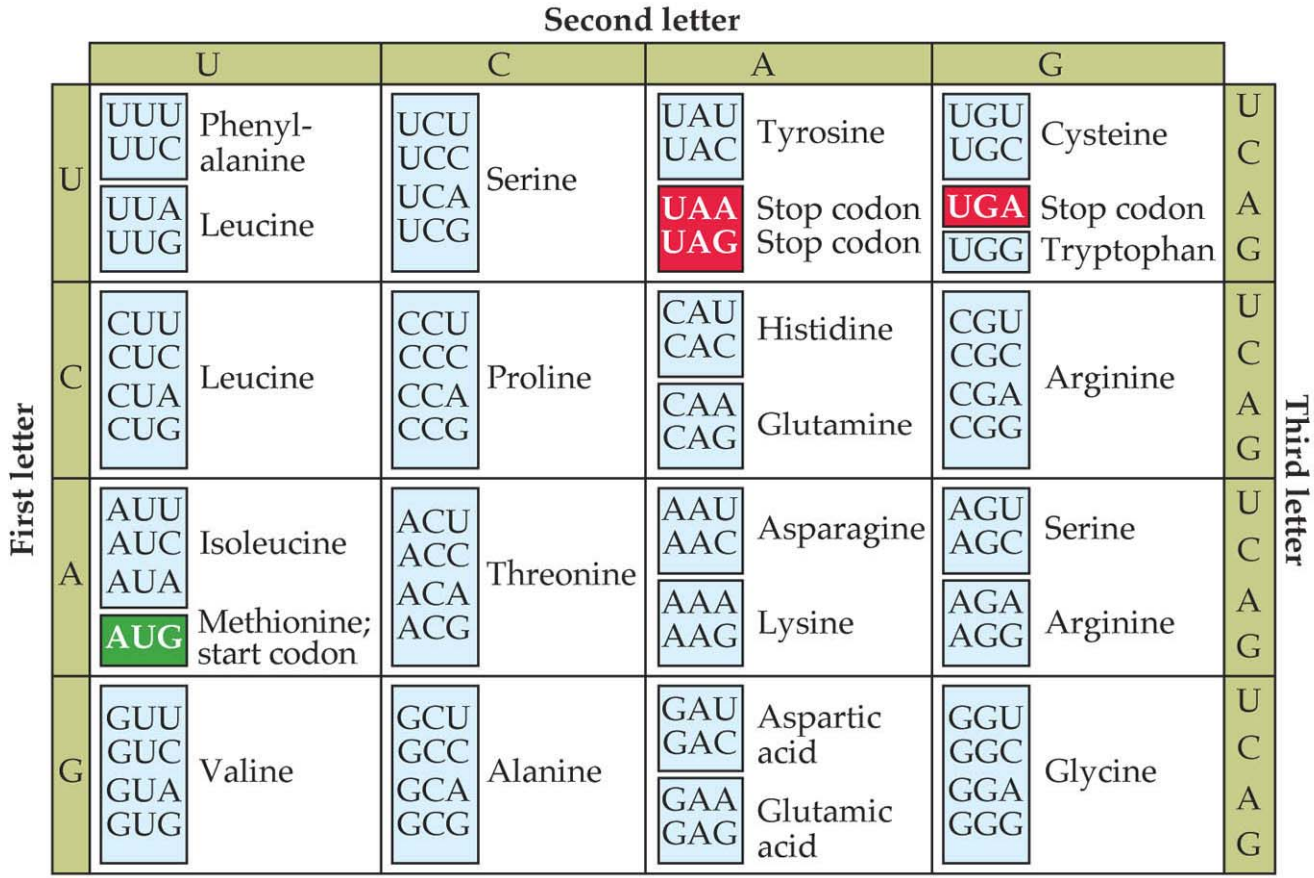
- *Potential:* Predicted using sequence analysis
- *Probable:* Uncertain experimental evidence
- *By similarity:* Predicted using sequence similarity

Cross references

Other databases (there are ~100 in total):

- Nucleotide sequences
- 3D structure
- Protein-protein interactions
- Enzymatic activities and pathways
- Gene expression (microarrays and 2D-PAGE)
- Ontologies
- Families and domains
- Organism specific databases

The genetic code



First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU, UUC: Phenylalanine	UCU, UCC, UCA, UCG: Serine	UAU, UAC: Tyrosine	UGU, UGC: Cysteine	U C A G
	UUA, UUG: Leucine		UAA: Stop codon; UAG: Stop codon	UGA: Stop codon; UGG: Tryptophan	
C	CUU, CUC, CUA, CUG: Leucine	CCU, CCC, CCA, CCG: Proline	CAU, CAC: Histidine	CGU, CGC, CGA, CGG: Arginine	U C A G
			CAA, CAG: Glutamine		
A	AUU, AUC, AUA: Isoleucine	ACU, ACC, ACA, ACG: Threonine	AAU, AAC: Asparagine	AGU, AGC: Serine	U C A G
	AUG: Methionine; start codon		AAA, AAG: Lysine	AGA, AGG: Arginine	
G	GUU, GUC, GUA, GUG: Valine	GCU, GCC, GCA, GCG: Alanine	GAU, GAC: Aspartic acid	GGU, GGC, GGA, GGG: Glycine	U C A G
			GAA, GAG: Glutamic acid		

- Degenerate (*redundant*) but not ambiguous
- *Almost* universal (deviations found in mitochondria)

Reading Frames 1

A piece of an mRNA-strand:

```
5'                                                              3'
augcccaagcugaauagcguagaggggguuuucaucauuugaggacgauguauaa
```

can be divided into triplets (*codons*) in three ways:

```
1 aug ccc aag cug aau agc gua gag ggg uuu uca uca uuu gag gac gau gua uaa
   M   P   K   L   N   S   V   E   G   F   S   S   F   E   D   D   V   *
2  ugc cca agc uga aua gcg uag agg ggu uuu cau cau uug agg acg aug uau
    C   P   S   *   I   A   *   R   G   F   H   H   L   R   T   M   Y
3   gcc caa gcu gaa uag cgu aga ggg guu uuc auc auu uga gga cga ugu aua
     A   Q   A   E   *   R   R   G   V   F   I   I   *   G   R   C   I
```

Each possible set of triplets is called a *reading frame.*

Reading Frames 2

Since there are two strands in DNA, there are *six* possible reading frames in a piece of DNA (three in each direction):

```
 3    A  Q  A  E  *  R  R  G  V  F  I  I  *  G  R  C  I
 2   C  P  S  *  I  A  *  R  G  F  H  H  L  R  T  M  Y
 1  M  P  K  L  N  S  V  E  G  F  S  S  F  E  D  D  V  *
5' ATGCCCAAGCTGAATAGCGTAGAGGGGTTTTCATCATTTGAGGACGATGTATAA 3'
3' TACGGGTTCGACTTATCGCATCTCCCCAAAAGTAGTAAACTCCTGCTACATATT 5'
    H  G  L  Q  I  A  Y  L  P  K  *  *  K  L  V  I  Y  L  -1
      G  L  S  F  L  T  S  P  N  E  D  N  S  S  S  T  Y   -2
     A  W  A  S  Y  R  L  P  T  K  M  M  Q  P  R  H  I    -3
```

A reading frame from a start codon to the first stop codon is called an *open* reading frame (underlined above).